

# First steps in the logic-based assessment of post-composed phenotypic descriptions

E. Jiménez-Ruiz<sup>1,3</sup>, B. Cuenca Grau<sup>2</sup>, R. Berlanga<sup>1</sup>, and Dietrich Rebholz-Schuhmann<sup>3</sup>

<sup>1</sup> Universitat Jaume I de Castelló, Spain {ejimenez,berlanga}@uji.es

<sup>2</sup> University of Oxford, UK {berg}@comlab.ox.ac.uk

<sup>3</sup> European Bioinformatics Institute, Cambridge, UK {jimenez,rebholz}@ebi.ac.uk

**Abstract.** In this paper we present a preliminary logic-based evaluation of the integration of post-composed phenotypic descriptions with domain ontologies. The evaluation has been performed using a description logic reasoner together with scalable techniques: ontology modularization and approximations of the logical difference between ontologies.

## 1 Introduction

A phenotype is defined as a basic observable characteristic of an organism. Thus, a set of phenotypic descriptions may involve different domains and granularities ranging from molecular to organism level.

Phenotypic descriptions have been recently described by means of terminological resources, with the Human Phenotype Ontology (HPO) [1] being a prominent example. The HPO ontology represents a so-called pre-composed description: it does not provide explicit links between the phenotypic descriptions (e.g. increased calcium concentration in blood) and the relevant entities associated to it, such as the chemical element involved (“calcium”), the way in which it is involved (“increased concentration”) and where it appears (“blood”). Post-composed phenotypic descriptions intend to provide a more formal representation to interoperate with involved entities [2] and to allow more powerful reasoning. Nevertheless, the formal representation of phenotypic descriptions is still a challenge [3, 4] owing to the complex nature of some phenotypes and the lack of consensus among clinicians to describe them in a standard way.

Mungall et al. [3] and Hoehndorf et al. [4] have recently proposed automatic and semi-automatic methods to transform pre-composed phenotypic descriptions into a *description logic* (DL) based post-composed representation linked to domain ontologies. The integration of domain ontologies with post-composed phenotypic descriptions presents new challenges since most of the involved ontologies are developed independently and may perform a different conceptualization for the same entities. Therefore, this integration may not always lead to the expected and proper logical consequences [5, 6]. In this paper we present first steps towards the logic-based assessment of the integration of phenotypic descriptions with domain ontologies.

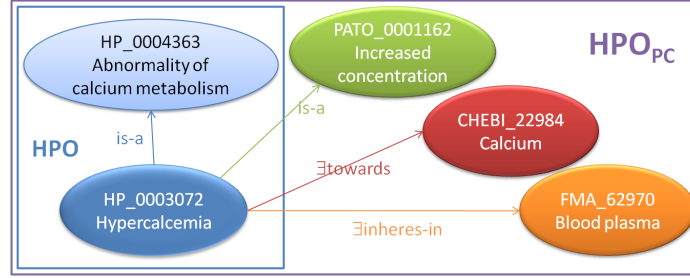


Fig. 1. An excerpt from the post-composed phenotypic descriptions of  $HPO_{PC}$

## 2 Method and preliminary results

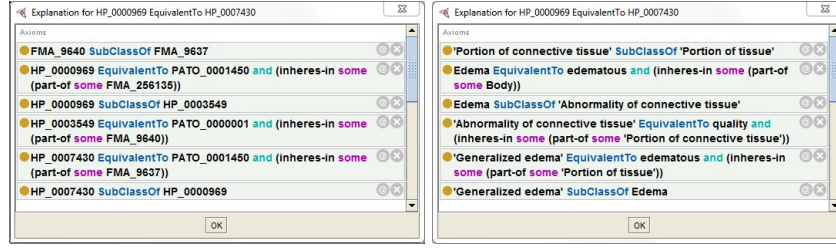
Our experiments have been based on a post-composed version (from now on  $HPO_{PC}$ ) of the HPO ontology<sup>4</sup> applying the method from [3]. The HPO ontology only provides a classification of pre-composed phenotypic descriptions (e.g. see left hand side of Figure 1), whereas  $HPO_{PC}$  also provides explicit links to relevant domain entities (see right hand side of Figure 1).  $HPO_{PC}$  contains 11382 entities and uses *external* concepts from different domain ontologies, including PATO [7] (264 concepts), Cell Ontology (12 conc.), GO (96 conc.), FMA [8] (812 conc.), CHEBI (33 conc.), and other OBO foundry ontologies [9].

A DL reasoner may be used to reclassify HPO concepts, according to the knowledge of  $HPO_{PC}$  and linked ontologies, and get new interesting knowledge. However, as stated in [3], reasoning with  $HPO_{PC}$  and all linked ontologies is time consuming. To smooth this limitation, we have extracted a locality-based module [10] for each set of referenced external entities. For example, the module for FMA contains 2044 concepts, which is much easier to reason with than the whole FMA (around 80000 concepts). Thus, we have built  $HPO_{ALL}$ <sup>5</sup> by merging  $HPO_{PC}$  with the corresponding modules from the referenced ontologies. The classification of  $HPO_{ALL}$  using HermiT [12] takes around 45 seconds in a 2Gb laptop.

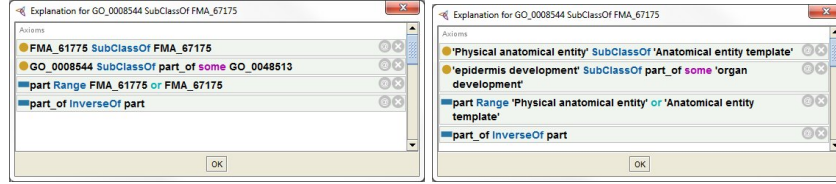
New subsumption relationships between HPO concepts may represent both desired new knowledge and unintended consequences. In order to evaluate the new logical consequences hold in  $HPO_{ALL}$  we have borrowed the notion of logical difference from [13]. The logical difference between two ontologies contains the set of consequences that are inferred in one of the ontologies but not in the other. Unfortunately, there is no algorithm for computing the logical difference in expressive DLs. Moreover, the number of inferences in the difference may be infinite. Thus, we have reused the approximations of the logical difference presented in previous work [5], where inferences are one of the following simple kinds of axiom: (i)  $A \sqsubseteq B$ , (ii)  $A \sqsubseteq \neg B$ , (iii)  $A \sqsubseteq \exists R.B$ , (iv)  $A \sqsubseteq \forall R.B$ , and (v)  $R \sqsubseteq S$  ( $A, B$  are atomic concepts, including  $\top, \perp$ , and  $R, S$  atomic roles).

<sup>4</sup> Available from <http://bioonto.de/obo2owl/hpo-in-owl.owl>

<sup>5</sup> We have converted the OBO ontologies to OWL using the OWLDEF method [11], and we have normalized the involved concept and property URIs



**Fig. 2.** Explanation for new equivalence between concepts *HP-0000969* (*Edema*) and *HP-0007430* (*Generalized edema*). With concept IDs (left) and concept names (right).



**Fig. 3.** Explanation for new subsumption relationship between concepts *GO-0008544* (*Epidermis development*) and *FMA-67175* (*Anatomical entity*)

The logical difference between  $HPO_{ALL}$  and  $HPO_{PC}$ , affecting only HPO concepts, contains 759 new subsumption relationships (inferences of type *(i)*). The integration leads indeed to a reclassification of HPO concepts. For example,  $HPO_{ALL}$  infers the probably non-intended consequence *Generalized edema*  $\equiv$  *Edema* which was not hold in  $HPO_{PC}$ . As shown in the Protégé-like explanation from Figure 2 the new knowledge from FMA leads to this new consequence.

The logical difference also contains 80 new entailments that relate concepts from domain ontologies (i.e. new cross-references). For example, the GO concept *Epidermis development* is classified under the FMA concept *Anatomical entity*. This consequence is probably not intended and it is due to the definition of range axioms in FMA (see Figure 3) and the use of the property *part\_of* in different scopes (in FMA relates anatomical entities, whereas in GO biological processes). Additionally, if a greater approximation of the logic difference is considered (i.e. entailments of type *(ii)-(v)*) new consequences are also obtained (e.g.  $GO\_0030308 \sqsubseteq \exists negatively\_regulates.GO\_0040007$ , where  $GO\_0030308$  stands for *Negative regulation of cell growth* and  $GO\_0040007$  stands for *Growth*).

### 3 Conclusions and future work

The benefits of integrating phenotypic descriptions with domain ontologies have already discussed in the literature [2, 3, 4]. However, the consequences of the integration should be evaluated by domain experts in order to detect potential unintended consequences.

In this paper we have performed a preliminary evaluation<sup>6</sup> in which state of the art techniques (e.g. ontology reasoning, ontology modularization, logical difference) have been reused to extract the set of new consequences when integrating post-composed phenotypic descriptions, such as the provided by **HPO<sub>PC</sub>**, with domain ontologies. In a near future, we intend to develop a system to guide the expert in the detection and repair of unintended consequences such as in our previous tool **ContentMap** [5], in which we assessed the integration of ontologies through mappings.

Moreover, domain ontologies contains cross-references (i.e. mappings) which have not been considered for this preliminary assessment. These new correspondences will probably lead to new consequences that should be assessed. Thus, we also intend to adapt the techniques proposed in [6] to this new setting.

## References

- [1] Robinson, P., Mundlos, S.: The Human Phenotype Ontology. *Clinical Genetics* **77**(6) (2010) 525–534
- [2] Lussier, Y.A., Li, J.: Terminological mapping for high throughput comparative biology of phenotypes. In: *Pacific Symposium on Biocomputing*. (2004) 202–213
- [3] Mungall, C., Gkoutos, G., Smith, C., Haendel, M., Lewis, S., Ashburner, M.: Integrating phenotype ontologies across multiple species. *Genome Biology* **11**(1) (2010) R2
- [4] Hoehndorf, R., Oellrich, A., Rebholz-Schuhmann, D.: Interoperability between phenotype and anatomy ontologies. *Bioinformatics* (2010)
- [5] Jiménez-Ruiz, E., Cuenca Grau, B., Horrocks, I., Berlanga, R.: Ontology integration using mappings: Towards getting the right logical consequences. In: *European Semantic Web Conference*. Volume 5554 of LNCS. (2009) 173–187
- [6] Jiménez-Ruiz, E., Cuenca Grau, B., Horrocks, I., Berlanga, R.: Logic-based assessment of the compatibility of UMLS ontology sources. Accepted for publication in *Journal of Biomedical Semantics* (2010)
- [7] Gkoutos, G., Green, E., Mallon, A.M., Hancock, J., Davidson, D.: Using ontologies to describe mouse phenotypes. *Genome Biology* **6**(1) (2004) R8
- [8] Mejino Jr., J.L.V., Rosse, C.: Symbolic modeling of structural relationships in the foundational model of anatomy. In: *Proceedings of KR-MED*. (2004) 48–62
- [9] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., et al.: The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology* **25**(11) (2007) 1251–1255
- [10] Cuenca Grau, B., Horrocks, I., Kazakov, Y., Sattler, U.: Modular reuse of ontologies: Theory and practice. *J. of Artificial Intelligence Research* **31** (2008)
- [11] Hoehndorf, R., Oellrich, A., Dumontier, M., Kelso, J., Rebholz-Schuhmann, D., Herre, H.: Relations as patterns: bridging the gap between OBO and OWL. *BMC Bioinformatics* **11**(1) (2010) 441
- [12] Motik, B., Shearer, R., Horrocks, I.: Hypertableau Reasoning for Description Logics. *Journal of Artificial Intelligence Research* **36** (2009) 165–228
- [13] Konev, B., Walther, D., Wolter, F.: The logical difference problem for description logic terminologies. In: *IJCAR*. Volume 5195 of LNCS., Springer (2008) 259–274

---

<sup>6</sup> **HPO<sub>ALL</sub>** and related domain ontology modules are available at: <http://krono.act.uji.es/people/Ernesto/phenotypeassessment/>